frontiers in ICT

METHODS
published: 26 December 2016
doi: 10.3389/fict.2016.00032

Developing an Integrated Image Bank and Metadata for Large-scale Research in Cerebrovascular Disease: Our Experience from the Stroke Image Bank Project

*Samuel O. Danso[1], Dominic E. Job[1], David Rodriguez Gonzalez[1], David Alexander Dickie[1], Jeb Palmer[1], Jenny Ure[2], Philip M. Bath[3], Peter A. G. Sandercock[1] and Joanna M. Wardlaw[1]**

[1] Brain Research Imaging Centre (BRIC), Centre for Clinical Brain Sciences (CCBS), University of Edinburgh Medical School, Edinburgh, UK, [2] MRC Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK, [3] Stroke Trials Unit (STU), Division of Clinical Neuroscience (DCN), University of Nottingham, Nottingham, UK

OPEN ACCESS

Edited by:
David N. Kennedy, University of Massachusetts Medical School, USA

Reviewed by:
Paul Kwan, University of Tsukuba, Japan
K. C. Santosh, University of South Dakota, USA

***Correspondence:**
Joanna M. Wardlaw
joanna.wardlaw@ed.ac.uk

Specialty section:
This article was submitted to Computer Image Analysis, a section of the journal Frontiers in ICT

Received: *26 August 2016*
Accepted: *08 December 2016*
Published: *26 December 2016*

Citation:
Danso SO, Job DE, Gonzalez DR, Dickie DA, Palmer J, Ure J, Bath PM, Sandercock PAG and Wardlaw JM (2016) Developing an Integrated Image Bank and Metadata for Large-scale Research in Cerebrovascular Disease: Our Experience from the Stroke Image Bank Project. Front. ICT 3:32. doi: 10.3389/fict.2016.00032

A framework for building an infrastructure that semantically integrates, archives, and reuses data for various research purposes in human brain imaging remains critical. In particular, problems of aligning technical, clinical, and professional systems in order to facilitate data sharing are a recurring issue in brain imaging. However, large samples of well-characterized images with detailed metadata are increasingly needed. This paper outlines the experience of the NeuroGrid Stroke Exemplar and further work in the Brain Research Imaging Centre and Stroke Trials Unit in developing an infrastructure that facilitates the linkage, archiving, and reuse of imaging data from stroke patients for large-scale clinical and epidemiological studies. We examined data from 12 past stroke projects carried out over the past two decades in our center and two large trials with 329 centers. We assessed previously published schemas and those developed specifically for large multicentre ischemic and hemorrhagic stroke treatment trials. We then developed our own harmonized and integrated schema and database with a web-based interface system, Longitudinal Online Research and Imaging System (LORIS), aiming to be flexible and adaptable to future trials and observational studies. We then linked image and metadata from 3,079 patients acquired in stroke research in one center in a 14-year period (1996–2010) with prospective central hospital health statistics to obtain long-term follow-up. Our integrated database includes 3,079 subjects and over 550 federated and searchable data items including imaging details, medical history, and examination, stroke, and laboratory details, which map to large multicentre stroke trials with imaging data from over 10,000 patients from 30 countries. The central linkage identified 879 of 3,079 patients had died, 525 had recurrent strokes, and 291 developed dementia during up to a 19-year period (range = 0–19; median = 9.04; IQR = 12.17) of follow-up, demonstrating its utility. The core metadata schema has benefited from extensive development in large clinical trials. Further trials' data can now be added. It provides an opportunity to crosslink and reuse data for a range of large-scale stroke

brain imaging clinical and research purposes including developing data analytics models for research into common brain diseases and their consequences.

Keywords: multicenter imaging, heterogeneous data, metadata schema, ischemic and hemorrhagic, image bank, neuroimaging, data sharing, stroke

INTRODUCTION

There is a global drive to develop strategies and frameworks to facilitate archiving, sharing, and reuse of data obtained from original research projects in order to maximize the value of the data (Pilat and Fukasaku, 2007; Walport and Brest, 2011; Mennes et al., 2013; Ferguson et al., 2014; Poldrack and Gorgolewski, 2014). This involves developing the required infrastructure that aligns technical, clinical, and biomedical systems and semantically integrates data from multiple sources, archiving, and making it available to be reused. Such integration is particularly important when creating large datasets from smaller individual studies for use in large-scale image analysis projects, especially for stratified medicine and machine learning which require very large amounts of individualized subject-specific data. In spite of the significant progress made in several neuroimaging domains such as the Biomedical Informatics Research Network (Keator et al., 2008); LORIS (Das et al., 2012), XNAT Central (Marcus et al., 2007); the Alzheimer's Disease Neuroimaging Initiative (Jack et al., 2008); the Human Connectome Project (Van Essen et al., 2013); and the BRAINS project (Job et al., 2016), the problem remains partially solved particularly for neurological diseases such as stroke (Warach et al., 2016).

Stroke researchers have access to imaging and associated data from multiple sources, in many different formats and at different levels of granularity. However, despite stroke being one of the most advanced fields among common neurological diseases in terms of (a) having a standard outcome measure for trials [the modified Rankin Scale (Lees et al., 2012)] and (b) effective treatments and prevention (Lindley et al., 2015), in general, the data collection protocols lack widely used standards, vary considerably, without clearly published provenance information between and within studies, which has significantly impeded the utility of the data (Ferguson et al., 2014; Nichols et al., 2016). Meanwhile, there would be numerous benefits that can be derived from semantically integrated data for various endeavors. Specifically, trials of new treatments for stroke require imaging data as part of the patient assessment (Wintermark et al., 2013), but the sample size needs to be large enough to obtain reliable results, particularly where treatment effects are likely to be modest (Lindley et al., 2015): the ability to combine image as well as clinical data facilitates meta-analyses (Laird et al., 2011). Furthermore, a semantically integrated patient database could be an efficient and cost-effective way to obtain data from many different centers and many different countries in order to obtain the sample size required to be able to observe a statistically significant difference between the subtypes of stroke and other key clinical variables or treatment effects in observational studies or clinical trials (Poldrack and Gorgolewski, 2014). Additionally, an integrated image bank offers the potential for building data analytics models, which will offer researchers the opportunity to develop new insight and understanding (Gomez-Cabrero et al., 2014).

The paper details our experience on the NeuroGrid Stroke Exemplar (Wardlaw et al., 2007) and further work that was carried out at the Brain Research Imaging Centre (BRIC), University of Edinburgh in collaboration with Stroke Trials Unit, University of Nottingham. The aim of the project was to develop an infrastructure to facilitate linkage, archiving, and reuse of neuroimaging data from stroke patients for large-scale clinical trials, focused observational, mechanistic, and epidemiological studies. We outline the recurring challenges associated with integrating neuroimaging data from multiple sources. We then describe the approach employed to develop an integrated metadata and schema for ischemic and hemorrhagic stroke, as the first step toward integrating neuroimaging data that combines clinical, demographic, and treatment data from patients. We further describe how we developed an integrated schema and database with a web-based interface system, with the aim of being flexible and adaptable to future trials and observational studies. We finally demonstrate the utility of the schema by linking the images and data to prospective central hospital health statistics.

Recurring Issues in Integrating Neuroimaging Data from Multiple Sources

Integrating and sharing imaging and associated data across multiple studies requires shared understanding of the datasets within the domain. Data from patients with common neurological disorders such as stroke are collected increasingly from a growing range of imaging modalities, especially computerized tomography (CT) and magnetic resonance (MR) imaging, and both produce multiple types of images. Images from different sites reflect differences in the scanner manufacturer and models used, and calibrations employed (Warach et al., 2016), even when similar MR sequences are deployed, although frequently MR in stroke still omits key sequences such as T2* weighted or T1 weighted. **Figure 1** shows an example of four early ischemic signs commonly seen in stroke patients imaged soon after stroke, distilled from a large literature survey to represent common features and terminologies (Wardlaw and Mielke, 2005) and which can then be captured efficiently by expert scan readers, e.g., in multicenter clinical trials, providing a simplified shared naming convention for ischemic lesions that allows translation between research and clinical practice.

However, even in an apparently simple process such as plain CT bran scanning (the commonest method used in stroke), there is variability in image and associated clinical data acquisition, transfer and storage that reflects the complexity, and variability in clinical practice as well as those that exist in the structural

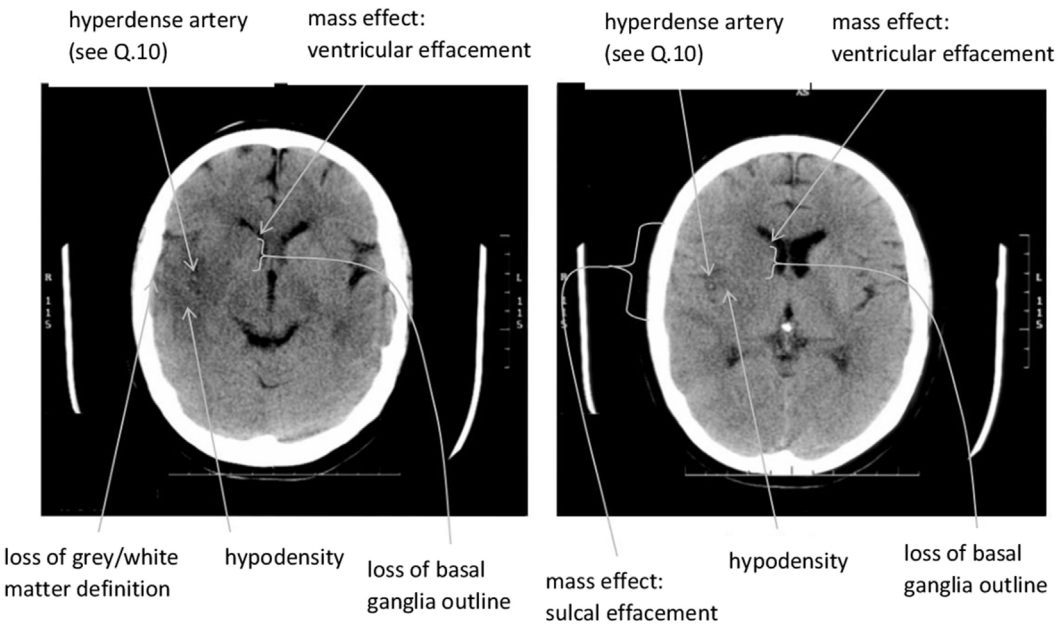

FIGURE 1 | Extract from scan reading pro forma in Third International Stroke Trial, illustrating a condensed and simplified terminology for the four features that are commonly seen in ischemic stroke. Developed from an extensive literature survey in Wardlaw and Mielke (2005) and other works. The features are (1) hypoattenuation (loss of gray/white differentiation or basal ganglia outline), (2) mass effect, (3) hyperdense artery (indicating thrombus), and (4) the lesion extent. Image © J. Wardlaw, reproduced with permission.

representation of the heterogeneous brain data (Keator et al., 2008). These issues have major integration challenges for machines (less so for humans), which can be addressed by metadata schema harmonization to achieve a simplified shared naming convention required in order to be accessible for machines (Keator et al., 2009). "Metadata" are facts about a given dataset that provides additional information regarding the parameters in which the dataset was acquired and the assumptions made about the experiment or analyses that helps one understand and use the data. For example, in the context of medical imaging data, metadata will allow machine-based reference models to be built and embedded into software for rapid determination of the validity of imaging data at the point of image acquisition. This is applicable to all data acquisition where imaging has a key role.

Progress toward Integrating Neuroimaging Data for Stroke Image Bank

Attempts are being made toward developing infrastructures to facilitate sharing and reuse of neuroimaging data from heterogeneous sources. To the best of our knowledge, **Table 1** shows all image banks specifically developed for stroke. We examine each briefly to determine their relevance and scope for stroke clinical trials.

The descriptions provided in **Table 1** demonstrate the scope and limitations of the existing stroke image banks, with respect to facilitating clinical trials of new treatments for stroke, which was the focus of the NeuroGrid project (Geddes et al., 2005; Wardlaw et al., 2007). NeuroGrid focused on two exemplar large multicenter clinical stroke trials that were ongoing at the time, the Third International Stroke Trial (IST-3) (Sandercock et al., 2012) and the Efficacy of Nitric Oxide in Stroke (ENOS) trial (The ENOS Trail Investigators, 2015). In order to create an integrated searchable database that could ultimately house the image data of both trials for future meta-analyses and data sharing to which other trials could be added, we had to design purpose-specific stroke imaging metadata and a related schema to accommodate different data structures and purposes, including, in addition to the actual images, collection of data on initial clinical assessments across several domains, long-term outcomes, treatments, and radiological interpretations of the images, which would be sufficiently flexible and adaptable for use in any future clinical trial or observational study in ischemic or hemorrhagic stroke (Wardlaw et al., 2007).

MATERIALS AND METHODS

The concepts and methods described here arose from NeuroGrid, followed by our work in developing an image bank of normal subjects across the lifespan in the BRAINS project[1] and also described in Job et al. (2016). The BRAINS project was carried out in parallel with adapting the stroke data schema to accommodate all data acquired in a series of 12 observational mechanistic and diagnostic studies in patients with various subtypes of stroke acquired in one center between 1996 and 2013 (but to which subsequent studies are being added).

[1]http://www.brainsimagebank.ac.uk.

TABLE 1 | Stroke image banks.

Reference	Stroke image bank project	Scope
Hanser et al. (2007)	neurIST	Focuses on very specific terminologies for describing vascular abnormality, clinical features, treatments and outcomes for subarachnoid hemorrhage
Colombo et al. (2010)	NeuroWeb	Focuses on genetics means that there is less priority given to recording image data in the detail required for many acute stroke treatment trials or other types of stroke research where highly specialized phenotyping including detailed imaging is required
Gibaud et al. (2011)	NeuroLOG	Focuses on the neuropsychological aspects of stroke and computational image analysis and does not provide for documenting more clinically relevant acute treatment and outcomes
Wang et al. (2011)	Medical Image Management System	This is particularly useful for managing imaging data in clinical trials but neither relevant to stroke specifically nor to observational studies with heterogeneous data
Ali et al. (2012)	Virtual International Stroke Trials Archive	Focuses on clinical stroke research for prevention, rehabilitation, imaging, and intracerebral hemorrhage. However, data are limited to demographic and clinical data from baseline and follow-up visits (2 h–90 days)
Wintermark et al. (2013)	Stroke Imaging Repository	Focuses on terminology and standardization for acute ischemic stroke trials but not metadata schema required for integrating heterogeneous imaging data (initiated with early terminology from NeuroGrid Stroke exemplar, an early version of the Stroke Schema in the present paper)
Kim et al. (2014)	CRCS-5	Focuses on ischemic stroke monitoring and management in hospitals. Also, although data are collected from multiple centers, it does not require metadata schema for integration as it uses a single data management with web-based interface system
Seghier et al. (2016)	PLORAS	Data are not heterogeneous and also focuses on only speech and language abilities-related outcomes of stroke

Our Approach

Image bank development begins with data integration. Data integration approaches could be broadly grouped into two. The "centralized approach" is where data sources are accessed through a single access point based on a predefined common metadata schema (Keator et al., 2009). The alternative is the "federation-based approach," which requires a framework in order to present a unified view of the data from multiple sources (Wiederhold, 1992). Our framework is, of necessity, federation-based, based on semantic rules derived from expert knowledge underpinned by many years of professional experience in stroke research including in clinical trials. **Figure 2** shows the schematic diagram of the framework, which we subsequently describe in detail.

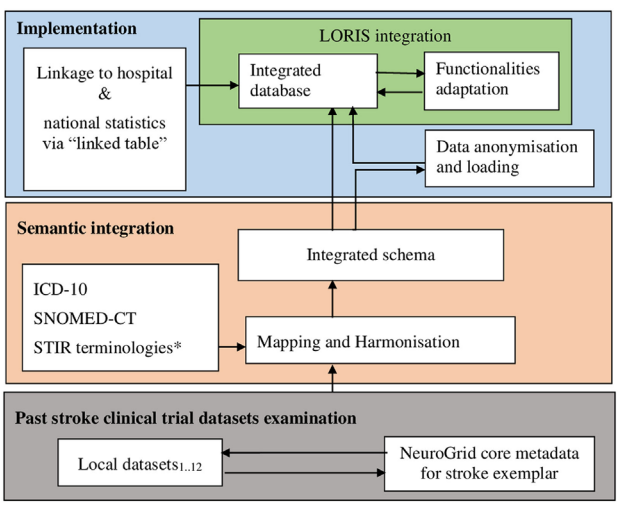

FIGURE 2 | Schematic diagram of the framework for the stroke image bank. LORIS, Longitudinal Online Research and Imaging System; ICD-10, the World Health Organization's International Classification of Diseases coding version 10; SNOMED-CT, a systematized nomenclature of medicine—clinical terms; STIR, Stroke Imaging Repository coding standards. *Initiated with terminology from NeuroGrid stroke exemplar, i.e., an early version of the present schema.

Step1: Examination of Datasets from Past Projects and NeuroGrid Stroke Example Metadata

As a first step toward developing an integrated schema, we started with the NeuroGrid schema based on the two large multicentre international trials, ENOS and IST-3, and examined data from 12 past stroke imaging research projects with various different objectives including different stroke subtypes and types of imaging, carried out over the past two decades in our center. These projects varied in research objectives and data collection protocols. This is demonstrated with two examples.

First, the Salvageable Tissue study (Wardlaw et al., 2013) was a multicenter study carried out in three acute stroke centers in Scotland (Aberdeen, Glasgow, and Edinburgh) between 2008 and 2010. The objective was to assess the practicalities of performing acute stroke imaging with CT and MR including perfusion imaging, to assess the proportion of patients with perfusion-evidence of salvageable tissue [perfusion-diffusion mismatch on MRI or reduced flow on CT perfusion (CTP)], and markers of

subsequent lesion growth on follow-up imaging to provide sample size estimates for future treatment trials. This involved recruiting patients with moderate to severe cortical ischemic stroke in three centers, performing imaging [diffusion weighted imaging (DWI), perfusion-weighted imaging, fluid attenuation inversion recovery (FLAIR), gradient echo (GRE/T2*), MR angiography (MRA); or with CT, CTP, and CT angiography (CTA)] within 6 h of stroke, repeated at 2–5 days (mostly MR) and 1 month (MR T2, GRE, DWI, and MRA). A final clinical follow-up was performed at 3 months.

The second is the Mild Stroke Study (Wardlaw et al., 2009) performed between 2005 and 2009. The aim was to investigate causes of lacunar stroke and associations with retinal vascular appearances (as a surrogate for cerebral small vessels). This was to test the theory that lacunar stroke and small vessel disease arise through blood–brain barrier damage. It recruited patients with lacunar or minor cortical ischemic stroke, all of whom had diagnostic MR imaging with DWI, FLAIR, T2-weighted, GRE, T1-weighted, and (in a subset) blood–brain barrier permeability imaging. A subset was followed up clinically and had follow-up imaging at 3 years after stroke.

The stroke exemplar metadata designed originally in the NeuroGrid project was an extension to the NeuroGrid core metadata and was designed to be scalable and modifiable to suit other stroke studies using imaging. The NeuroGrid core metadata was constructed to accommodate studies in stroke, dementia, and psychosis and was in response to one of the key infrastructure objectives of NeuroGrid—to develop management systems to allow large "living archives" of images linked to key metadata for diseases that require long-term study to understand their true natural history and the effects of treatment (Wardlaw et al., 2007). This involved developing a simple repository browser to perform *ad hoc* searches against the core metadata and display user-readable, navigable listings of search results including the images for administration and quality control. An example of a search could be to generate a list of all patients in trial X who were scanned at location Y and had a clinical feature Z and an imaging feature A.

In the stroke exemplar, the NeuroGrid core metadata schema was extended significantly based on the two large multicentre randomized stroke trials, IST-3 and ENOS. IST-3 was a 3035-patient multicenter randomized controlled trial of alteplase given up to 6 h after onset of acute ischemic stroke (Sandercock et al., 2008, 2012). IST-3 sought to determine whether a wider range of patients might benefit from intravenous recombinant tissue plasminogen activator (rt-PA). ENOS (The ENOS Trial Investigators, 2006, 2015) was a 4011-patient multicentre randomized controlled trial in patients with acute (<48 h of onset) ischemic or hemorrhagic stroke. ENOS tested the safety and efficacy of transdermal GTN, and of continuing or stopping temporarily prior antihypertensive medication. Both the trials required a CT brain scan at randomization (minimum requirement plain non-contrast CT brain), but MRI could be used instead (minimum sequences T2-weighted, FLAIR, DWI, and GRE). Advanced imaging, such as CTA, MRA, or perfusion imaging, was also collected where performed. Both the trials involved multiple centers ($n = 329$), and therefore, inevitably the images came from a very large variety of scanners (Wardlaw et al., 2007).

The extension of the core metadata schema was governed by issues relating to where, when, and how datasets are collected, published to the database, or required by clinicians. Thus, the resulting extended NeuroGrid core metadata for stroke allowed a search across a wide range of patient baseline characteristics (including history factors: vascular risk factors, prior treatments, past medical history), stroke clinical characteristics (severity, clinical subtype, neurological examination details), type and timing of imaging, appearance of the stroke lesion on imaging (including site and size), laboratory test results, details of trial treatment administration, details of any non-trial treatments, subacute and late clinical functional measures (symptomatic intracranial hemorrhage or brain swelling, modified Rankin Scale, death), cognitive and imaging outcomes, and adverse events.

We then compared our 12 study datasets from our center with the NeuroGrid stroke exemplar metadata. We noted the differences and overlaps that existed and iterated modifications to address items that were not covered in the original NeuroGrid exemplar or that were present but required more granularity and fed this into the subsequent developments of the data schema. We demonstrate this with some examples of the differences that were observed in data collection protocols between the Salvageable Tissue and Mild Stroke Studies described earlier. For example, the NeuroGrid exemplar schema required information about stroke severity using the National Institute of Health Stroke Scale (NIHSS) (Goldstein et al., 1989). While the Salvageable Tissue protocol required a detailed data to be recorded for each symptom (e.g., "Bast gaze," which is one of the items on the NIHSS is recorded as either "forced deviation" or "Normal" or "Partial gaze palsy"), the Mild Stroke Study protocol, on the other hand, required summary data, which is the total score assigned to each NIHSS symptom to be recorded. The reverse of this was observed in another instance. The NeuroGrid exemplar schema required data on classification of stroke based on the Oxford Community Stroke Project classification—OCSP (Bamford et al., 1987). In this instance, The Salvageable Tissue protocol required a summary of the data by recording either "present" or "not present" for each of the classifications [e.g., Partial Anterior Circulation Syndrome (PACS) is to be recorded as either "present" or "not present"] based on the assessment and knowledge of the clinician. On the other hand, the Mild Stroke Study protocol did not rely on the knowledge of the clinician to classify but only required data to be collected on symptoms such as weakness/sensory deficit in arm, leg, and face. The differences in data as result of differences in collection protocols demand some amount of adaptation from data integration and image bank perspective, which is subsequently described in step 2. The guiding principles adopted in this work were that the approach must be pragmatic; the metadata and schema should be relevant to clinical practice, as well as scalable to other researches where details might need to be added or switched off in particular domains, without requiring major redesign.

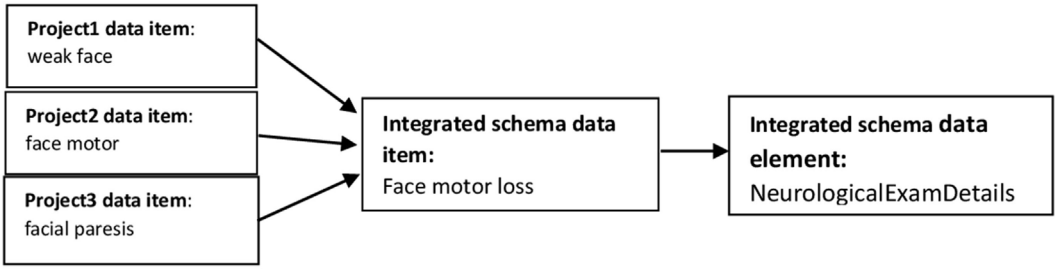

FIGURE 3 | **Mapping "face motor loss" variables as expressed in various datasets**.

Step 2: Semantic Integration

"Semantic integration" is the process of ensuring that all semantically related data elements and items are grouped together based on expert knowledge of domains and other resources. This was achieved through a series of steps described below.

Mapping and Harmonization

Mapping ensures that data items that have different names, but that are considered to be semantically the same or very similar, are captured as a single schema data item. This involved mapping the IST-3 and ENOS trials metadata and schema developed in NeuroGrid, then refining, and extending the schema based on the process described in step 1 above. Examination of the 12 local prior stroke research projects showed a high degree of variability in the datasets (from the machine point of view though not the human point of view), which is noted to be a common issue associated with data from multiple sources (Gomez-Cabrero et al., 2014), or in this case, even from a series of studies of one disease in one center that basically collected the same clinical variables even though each study might collect some other information. **Figure 3** illustrates an example of the variabilities and how these are handled.

For example, **Figure 3** shows three different variables ("weak face," "face motor," and "facial paresis") in three different projects being mapped to a single search item "face motor loss," which is part of the integrated schema data element, "NeurologicalExamDetails." On the other hand, harmonization is a process that ensures uniformity in how schema search items are encoded and represented. For example, "lesion age" in one dataset is encoded in categories (1 = "less than 6 h"; 2 = "6–12 h"; 3 = "greater than 12 h"), whereas in another dataset, different encoding scheme (e.g., raw values) are employed. Specifically, with regards to the examples of the problems between the Salvageable Tissue study and Mild stoke dataset described in step 1 above, the data on the individual symptoms were mapped to the corresponding numeric values for each symptom based on the NIHSS documentation (Goldstein et al., 1989). This enabled us to transform the responses into a total score representing the severity of stroke for each patient as required by our new metadata schema. Again, to be able to harmonize the OSCP data, rules were developed to transform the symptoms collected by the Mild stroke study based on the OSCP classification rules. So for example, if a patient had weakness and/or sensory problems in the face, arm, or leg and also has dysphasia, the stroke is classified as PACS being "present," otherwise "not present." Thus, reasonable encoding and representation were achieved through harmonization. This strategy was applied to all issues that were identified and documented as part of the provenance, which is also made available to potential users of the image bank. This process was automated using the Python programming language (version 3.2, see Python Software Foundation[2]).

Use of Coding Standards

In order to further enhance the interoperability and reusability of the integrated schema and image bank to facilitate future integration with other biomedical ontologies, we cross compared our terms with other data coding standards and medical taxonomies. This included standard terminologies that were originally derived from the NeuroGrid work with additional modification for use in the Stroke Imaging Repository of acute treatment and secondary prevention stroke trials (Wintermark et al., 2013), which also aligns with the National Institute of Neurological Disorders and Stroke Common Data Elements.[3] The World Health Organization's International Classification of Diseases coding version 10[4] and the systematized nomenclature of medicine—clinical terms (SNOMED-CT) (Cote and Robboy, 1980) provide a familiar and useful common vocabulary in clinical practice where other relevant data may be cross-referenced. ICD-10 and SNOMED-CT, in particular, are implemented as standards by health services in many countries hosting multi-site trials and has the additional benefit that allows integration with national health information systems and electronic health records (Westra et al., 2015).

Figure 4 shows schematic diagram of the integrated metadata schema with its data elements, which have over 550 integrated searchable data items contained within them.

As demonstrated in **Figure 4**, the resulting integrated schema will allow searches across a wide range of patient baseline and outcome characteristics described as part of the stroke exemplar and additional searchable data elements and items including read-by-an-expert, visual scores, and computationally measured imaging features. This includes categorization of the acute stroke

[2]https://www.python.org/.

[3]http://www.ninds.nih.gov/research/clinical_research/toolkit/common_data_elements.htm.

[4]http://apps.who.int/classifications/icd10/browse/2016/en.

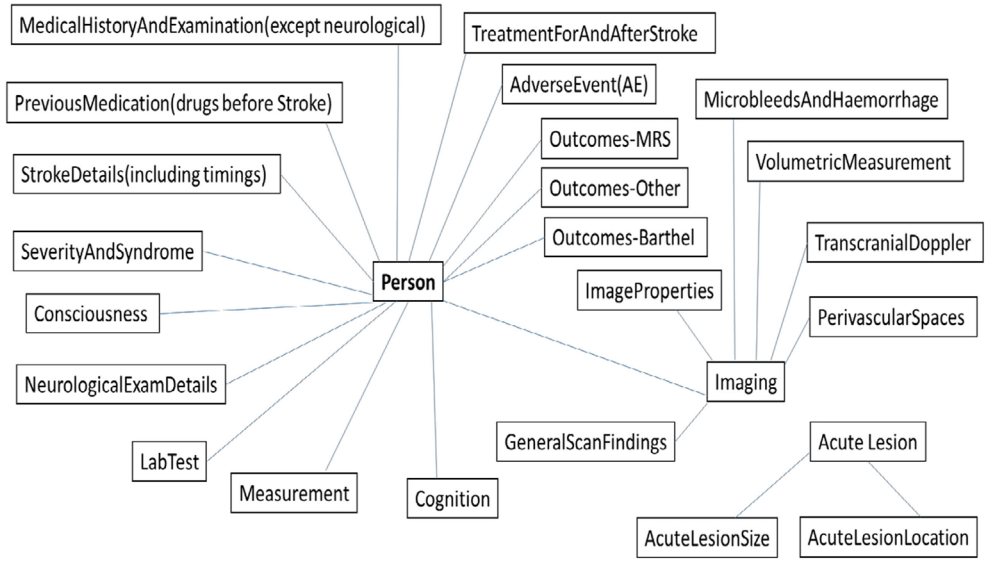

FIGURE 4 | Overview of our integrated schema for the stroke image bank with each box indicating an "element," or domain of information, and each element containing searchable "items" in the database. Outcomes-MRS: modified Rankin Scale; Outcomes-Other: the Nottingham Extended Activities of Daily Living Scale, Quality of Life, Disposition, and Death; Outcomes-Barthel: Barthel or Barthel Activities Daily Living Scale; Stroke Details (including timings): types of stroke.

lesion (infarct or hemorrhage, extent, background brain changes); volumetric measurements (e.g., intracranial volume, brain volume, infarct volume, white matter hyperintensity volume); other visual scores as relevant to, for example, small vessel stroke (e.g., perivascular spaces, lacunes, microbleeds by brain region); and lesion-specific anatomical locations (e.g., thalamus, gray white matter, deep white matter) where relevant.

Step 3: Implementation

Our implementation took advantage of available open source technologies as described below.

Longitudinal Online Research and Imaging System (LORIS) Integration

We integrated our integrated schema with the Longitudinal Online Research and Imaging System (LORIS) database in order to take advantage of its capabilities. LORIS is an open-source data management system, well engineered for managing imaging and associated behavioral longitudinal data, and implemented using MySQL and NoSQL (CouchDB)[5] for back-end web interface and Hypertext Preprocessor (PHP) programming language[6] for front-end web interface (Das et al., 2012), which we deployed in Linux Ubuntu 14.04 box.

Our clinical trial datasets also have longitudinal characteristics as projects required subjects to be followed up after the initial visit, sometimes over many years. Therefore, it was prudent to take advantage of the functionalities available in LORIS in order to avoid duplication of effort. MySQL, NoSQL and PHP are both open source and widely used relational database management systems and frameworks (Bakken et al., 1997; Bretthauer, 2002). Both MySQL and NoSQL as employed in LORIS offered us the following database design capabilities: (a) performance, which was to ensure speed processing of queries and a quick access to the data; (b) integrity, which was to ensure accurate storage of the data as obtained from the original sources; (c) comprehensibility, which was concerned with ensuring coherence in the structure of the database as presented to users; and (d) extensibility, which was to ensure the database can be extended without the need to redesign. The functionalities adaptation process involved integrating our Python-based scripts with the PHP-based script functionalities used in LORIS. The integration process was achieved through collaboration and support from the LORIS software development team.[7]

Data Anonymization and Loading

All images had already been anonymized of metadata by passing through DICOM Confidential (González et al., 2010), a freely available data anonymization tool for imaging.[8] It is a Java-based de-identification toolkit that enforces confidentiality policies as defined by the Medical Research Council.[9] It is also specifically designed to support batch processing for multicentre clinical trials. Additionally, all identifiable information contained in the columns of the associated clinical data was also removed to ensure complete anonymity. After the data anonymization process, we

[5]http://couchdb.apache.org/.
[6]http://php.net/manual/en/intro-whatis.php.
[7]http://loris.ca/.
[8]https://sourceforge.net/projects/privacyguard/.
[9]https://www.mrc.ac.uk/documents/pdf/personal-information-in-medical-research.

Stage	Status	Date
Screening		
Visit	Pass	2010-05-19
Approval	Pass	

Details of Imaging Performed

Date of Administration	2010May19
Candidate Age (Months)	740
Window Difference (+/- Days)	N/A
Examiner	
Date MRI was done	2010-05-20
MRI Diffusion weighted imaging	yes
MRI T2 imaging	yes
MRI T1 imaging	yes
MRI FLAIR imaging	yes
MRI Susceptibility weighted imaging	yes
MRI Gradient Echo imaging	no
MRI Spectroscopy imaging	no
MRI Spectroscopy imaging type	not_answered
MRI Chemical shift imaging	no
Other MRI imaging	not_answered
Date CT was done	not_answered
CT Plain	not_answered
CT Perfusion imaging	not_answered
Date Transcranial Doppler Ultrasound (TCD) was done	not_answered

QC Status	PSCID	DCCID	Visit Label	Site	QC Pending	DOB	Gender	Output Type	Scanner	Subproject
	MSSB009	836993	V0	Edinburgh Imaging (BRIC)		1935-03-17	Female	native	GE MEDICAL SYSTEMS Signa HDxt 00000000200MRS03	MSS2

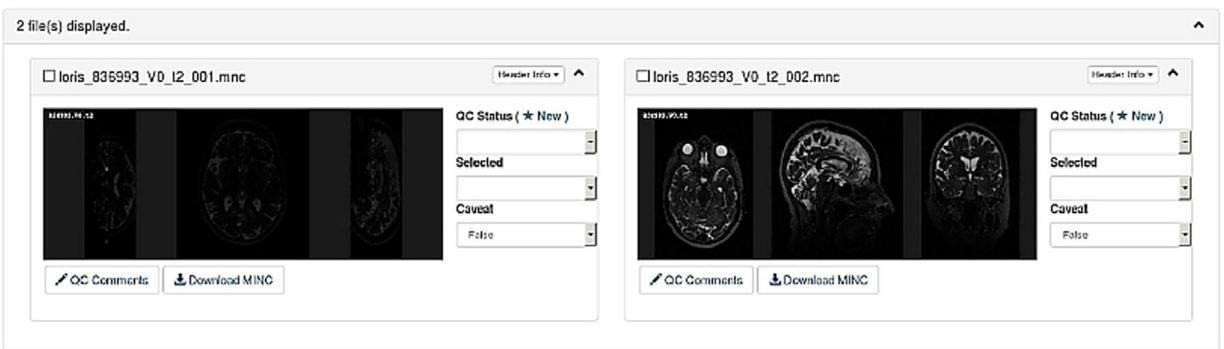

FIGURE 5 | A LORIS-based web interface of the stroke image bank showing details of an anonymized patient's imaging and clinical data as contained in the integrated database at baseline visit (visit 0).

then loaded the data by populating the integrated database with data from the clinical the trial datasets described in step 1 above. The loading process also accounts for the mapping and harmonization process that was carried out to ensure that the correct data items were populated to conform to our new integrated schema. This process was also automated using Python-based scripts.

Linkage to Hospital and National Statistics

We made provision for linking the integrated imaging database to hospital and national statistics to obtain long-term outcomes such as recurrent stroke, dementia, other vascular events, and death. We first obtained regulatory approvals from the relevant institutions. This include Caldicott Guardian and Community Health Index Advisory Board, NHS Lothian (reference: CG/DF/1559); NHS Lothian Research & Development (reference: 2015/0296); Information Services Division (ISD) and Scottish Stroke Care Audit (reference eDRIS-1516-0337); and West of Scotland Research Ethics Service (reference: 15/WS/0157). This allowed us to create a database of identifiable details of subjects scanned at our center in Edinburgh for the purpose of central matching with routinely collected health data by the Information Services Division of NHS Scotland.[10] In order to achieve the linkage between our integrated database and hospital and national statistics database, a "linked table" was created which holds the patients' hospital primary IDs and randomly generated IDs assigned to subjects in the integrated database by LORIS-based ID generation algorithm. Access to the linked table is restricted and only accessible to key approved members of research team covered by the data access agreements. The data anonymization and loading step described above also populated the integrated database with the individual "key" stored in the linked table.

Quality Control

In order to ensure data accuracy and consistency, an end-to-end quality control procedure was performed on samples of the data. This involved randomly selecting sample records from the web interface and checking data values against the source as well as data provenance.

RESULTS

Our integrated schema contains over 550 searchable data variables. Additionally, the integrated schema maps to IST3[11] and ENOS,[12] which are the two original NeuroGrid exemplar large multicentre stroke trials with over 7,000 patients from 30 countries between them. This demonstrates its utility within the context of ensuring data standards to facilitate seamless integration of heterogeneous multicentre neuroimaging data for ischemic and hemorrhagic stroke as well as stroke subtypes such as small vessel lacunar stroke. Moreover, our integrated database contains over 3,079 unique subjects from our 12 research studies, who were scanned in our local BRIC, Edinburgh, with neuroimaging data for ischemic and hemorrhagic stroke and small vessel disease studies. **Figure 5** shows the LORIS-based interface of our integrated database.

We submitted records on 3,245 patients from the combined dataset of 12 stroke studies in our 1 center for central linkage with routinely collected health records achieving an overall linkage success rate of 95% with the National Health Service (NHS) Hospital Information System and Stroke Audit databases of Scotland. A detailed breakdown showed that up to 19 years since inclusion in the research project and scanning (median = 9.04; IQR = 12.17, range 0–19 years) of follow-up, 879/3079 patients had died, 525 had had one or more recurrent stroke, and 291 had developed dementia, which further demonstrates the utility of our integrated database. The metadata schema for the integrated database and provenance information including data dictionary are available online under Apache 2.0 and CC-YB 4.0 licenses, respectively.[13]

DISCUSSION

Our neuroimaging data acquisition and management for stroke research has evolved from large pragmatic clinical stroke trials of acute stroke treatments with fairly basic imaging in NeuroGrid in the mid-2000s to include much more detailed bespoke observational mechanistic studies with much more complex imaging and longer follow-up linked with more detailed outcomes. This evolution demanded new approaches and also presents new opportunities. With the advent of "big data" science for medical and clinical research (Wang and Krishnan, 2014) and also for neuroimaging (Van Horn and Toga, 2014), our image bank will provide stroke researchers with new opportunities to explore big data science for stroke. An image bank with special focus on ischemic and hemorrhagic stroke and subtypes such as small vessel disease adds substantially to the dynamic range of capabilities of secondary research with cerebrovascular diseases data, thereby contributing to the volume and veracity of stroke data which characterize big data (Laney, 2001). Furthermore, employing international data standards facilitates the creation of Linked Data (Heath and Bizer, 2011), thus expanding the data space useful for new data management and technological initiatives for stroke. Also, the provision made in our integrated database to allow data from hospital information systems and national statistics to be linked provides opportunities to investigate a range of clinically highly relevant issues in stroke and to make use of centrally housed routinely collected image data in National Picture Archiving and Communication Systems PACS, such as the many thousands of brain scans collected in the first 8 years of the Scottish National PACS, now stored at the Farr Institute, Edinburgh.[14] To demonstrate this potential, for example, we are currently using imaging data from our 12 stroke studies linked to data from NHS Scotland's Information System and Stroke Audit databases to investigate imaging predictors of neurodegeneration measured at presentation with suspected stroke and subsequent adverse outcomes of recurrent stroke, dementia, or death.

From image analysis perspective, well-characterized images with detailed metadata are increasingly needed for studies that typically need larger samples or more variety of cases than are available in individual studies—these include studies to develop machine learning methods for image analysis, in stratified medicine, and large studies of genetics, e.g., genome wide association

[10] http://www.isdscotland.org/.

[11] http://www.dcn.ed.ac.uk/ist3/.

[12] http://www.strokecenter.org/trials/clinicalstudies/the-efficacy-of-nitric-oxide-in-stroke-enos-trial.

[13] https://sourceforge.net/projects/cvd-db.brainsimagebank.p/.

[14] http://www.farrinstitute.org/.

studies where typically many thousands of cases are needed (Hernández et al., 2013; Caligiuri et al., 2015). The availability of large amount of data could help develop models that can be generalizable based on the patterns the underlying algorithms are able to "learn" from the data. Large amounts of data can also provide enough statistical power for valid conclusions to be drawn (Cooper et al., 2011). This could be achieved by having access to selected cases with particular characteristics that are pulled from multiple studies for testing these algorithms and hypothesis. For example, Maillard et al. (2008) demonstrated the usefulness of image bank when they pulled over 1,100 of elderly subjects (with similar characteristics) from two large MRI studies to evaluate the performance of an automated method for detection, quantification, localization, and statistical mapping of white matter hyperintensities in T2-weighted images. An integrated image bank such as this will afford researchers the opportunity to carry out similar studies.

The framework that we employed offers an alternative to other frameworks proposed in the literature. The ontology-based federation is the most common approach within the neuroimaging domain (Hanser et al., 2007; Colombo et al., 2010; Gibaud et al., 2011). These approaches tend to rely on some specialized ontology to serve as a mediation layer between databases to integrate heterogeneous neuroimaging datasets (Wiederhold, 1992) and require that all potential submitters of data to the database stick religiously to the described schema terminology, which in reality is difficult across multiple sites. Within the context of stroke, the neurIST Project employed description logic-based ontology to represent concepts that are associated with cerebral aneurysms and subarachnoid bleedings (Hanser et al., 2007). Similarly, an ontology-based approach was also employed in the NeuroLOG (Gibaud et al., 2011) as well as NeuroWeb (Colombo et al., 2010) projects. A hybrid approach has also been proposed by Keator et al. (2013), where an ontology-based resource, NeuroLex (Larson and Martone, 2013), is combined with information obtained from other resources such as the Human Imaging Database[15] and XNAT.[16] None of these were suitable for stroke, thereby suggesting that lack of ontology for a given specialized domain raises significant neuroimaging data integration challenges (Smith et al., 2015). Furthermore, it has been noted that ontology-based approaches result in tensions between logical (research) and clinical representations of a domain, which make it difficult to create shared models resulting in tensions between ontological consistency and clinical usability (Bodenreider, 2004; Bodenreider and Stevens, 2006; Rector and Rogers, 2006). Thus, our approach is an important advance that overcomes the lack of a specialized ontology for ischemic and hemorrhagic stroke.

Moreover, there is an implicit expectation that medical concepts of disease, based on signs and symptoms, can be transposed as formally defined classes and relations, which are often much more complex to model in practice and resistant to simplification. Thus, the pragmatic and simplified approach adopted here makes our framework and data integration approach easy to implement. However, it is important to note that this is heavily dependent for its development on domain knowledge. In our case, the domain experts lead the project and were motivated to combine their datasets from individual studies, thus providing the required domain and semantic knowledge. Such exercises are not achievable without the close working of experts in the disease of interest (and in this case its imaging) with experts in the technological infrastructure required to host complex interrelated medical and imaging data, the former having the motivation and the content knowledge and the latter the essential knowledge to manage the data efficiently.

The mapping and harmonization process described as part of our framework involved data provenance documentation of the integrated schema.[17] This provides a detailed account of processes carried out on the datasets from the point of acquisition, descriptions of the imaging hardware and parameters used in the acquisition of the data, as well as mapping and harmonization (including transformations) as previously described (MacKenzie-Graham et al., 2008). The importance of this information has been emphasized (Keator et al., 2013) and documented as one of the guiding principles of data sharing best practices (Nichols et al., 2016).

CONCLUSION

This paper summarizes our experience in developing an integrated image bank and schema suitable for hosting data from multiple individual stroke imaging research projects and enabling large-scale research in cerebrovascular diseases, with a particular focus on ischemic and hemorrhagic stroke and small vessel diseases. This will facilitate research into new treatments for stroke by enabling large meta-analysis as well as testing computationally based image analysis methods (e.g., machine learning) for building predictive models specifically for stroke and other related conditions. In addition to adding more research data, we open the door to adding new data such as that routinely collected in health services, for example, by using Natural Language Processing (Chapman et al., 2011). Additionally, the past decade has seen unprecedented attempts to develop frameworks and infrastructure that can facilitate integration, archiving, and reuse of neuroimaging from multiple sources. We believe that the experience and framework described in this manuscript could be applied to neuroimaging data from other domains where resources such as ontologies do not currently exist.

AUTHOR CONTRIBUTIONS

JW, DJ, JP, JU, PB, and PS designed and carried out the NeuroGrid Stroke Exemplar project. JW, SD, DJ, DG, and DD created the integrated stroke metadata schema and image bank. All the authors contributed to the drafting of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank a wide range of colleagues in NeuroGrid and other HealthGrid projects who have

[15] http://www.nitrc.org/projects/hid/.

[16] http://www.xnat.org/.

[17] https://sourceforge.net/projects/cvd-db.brainsimagebank.p/.

contributed to the work described in this paper. Special thanks go to Andrew Duffy, the Farr Institute, and the NHS Scotland for extracting data for the image data bank. Finally, the authors are also thankful to Christine Rogers and the LORIS team at the McGill Centre for Integrative Neuroscience for their support in integrating LORIS with our federated database.

FUNDING

The Medical Research Council (Grant Ref no: GO600623 ID number 77729) for the NeuroGrid project. The SINAPSE Collaboration (Scottish Imaging Network, A Platform for Scientific Excellence, www.sinapse.ac.uk) through the Scottish Funding Council part funded JW. PB is Stroke Association Professor of Stroke Medicine and a NIHR Senior Investigator. This work was further supported by INNOVATE-UK (reference 102167) for the vascular linkage project. Also, as part of this work, SD received additional funding from Scottish Funding Council through the SINAPSE Postdoctoral and Early Career Researcher Exchange programme, which enabled him to visit Harvard Medical School, USA. ENOS was funded by the Bupa Foundation and Medical Research Council. The IST-3 trial was funded by the following agencies: the Australian Heart Foundation (Australian, grant number G 04S 1638); Australian NHMRC (grant number 457343); Danube University (Austria); the Dalhousie University Internal Medicine Research Fund (Canada); Norwegian Research Council (Norway); Polish Ministry of Science and Education (Poland, grant number 2PO5B10928); AFA Insurances (Sweden), the Swedish Heart Lung Fund (Sweden), Karolinska Institutet (Sweden), Stockholm County Council and Karolinska Institute Joint ALF-project grants (Sweden); Swiss National Science Foundation (Switzerland); Swiss Heart Foundation (Switzerland); The Foundation of Marianne and Marcus Wallenberg (Sweden); Foundation for health and cardio-/neurovascular research (Switzerland); the Medical Research Council (UK, grant numbers G0400069 and EME 09-800-15), The Health Foundation (UK), The Stroke Association (UK); DeSACC (UK); The University of Edinburgh (UK): The Lothian Health Board (UK); The Assessorato alla Sanita (Italy).

REFERENCES

Ali, M., Bath, P., Brady, M., Davis, S., Diener, H. C., Donnan, G., et al. (2012). Development, expansion, and use of a stroke clinical trials resource for novel exploratory analyses. *Int. J. Stroke* 7, 133–138. doi:10.1111/j.1747-4949.2011.00735.x

Bakken, S. S., Aulbach, A., Schmid, E., Winstead, J., Wilson, L. T., Lerdorf, R., et al. (1997). *PHP Manual*. Zend Technologies, Ltd. Available at: ftp://ftp.nyme.hu/doc/php/manual_m-x.pdf

Bamford, J., Sandercock, P., Jones, L., and Warlow, C. (1987). The natural history of lacunar infarction: the Oxfordshire Community Stroke Project. *Stroke* 18, 545–551. doi:10.1161/01.STR.18.3.545

Bodenreider, O. (2004). "The ontology-epistemology divide: a case study in medical terminology," in *Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004)* (Torino: IOS Press).

Bodenreider, O., and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Brief. Bioinformatics* 7, 256–274. doi:10.1093/bib/bbl027

Bretthauer, D. (2002). Open source software: a history. *Inform. Technol. Libr.* 21, 3–10.

Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. *Neuroinformatics* 13, 261–276. doi:10.1007/s12021-015-9260-y

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J. Am. Med. Inform. Assoc.* 18, 540–543. doi:10.1136/amiajnl-2011-000465

Colombo, G., Merico, D., Boncoraglio, G., De Paoli, F., Ellul, J., Frisoni, G., et al. (2010). An ontological modeling approach to cerebrovascular disease studies: the NEUROWEB case. *J. Biomed. Inform.* 43, 469–484. doi:10.1016/j.jbi.2009.12.005

Cooper, R., Hardy, R., Sayer, A. A., Ben-Shlomo, Y., Birnie, K., Cooper, C., et al. (2011). Age and gender differences in physical capability levels from mid-life onwards: the harmonisation and meta-analysis of data from eight UK cohort studies. *PLoS ONE* 6:e27899. doi:10.1371/journal.pone.0027899

Cote, R. A., and Robboy, S. (1980). Progress in medical information management: systematized nomenclature of medicine (SNOMED). *Jama* 243, 756–762. doi:10.1001/jama.1980.03300340032015

Das, S., Zijdenbos, A. P., Vins, D., Harlap, J., and Evans, A. C. (2012). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinformatics* 5:37. doi:10.3389/fninf.2011.00037

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat. Neurosci.* 17, 1442–1447. doi:10.1038/nn.3838

Geddes, J., Lloyd, S., Simpson, A., Rossor, M., Fox, N., Hill, D., et al. (2005). "NeuroGrid: collaborative neuroscience via grid computing," in *Proceedings of All Hands Meeting*. Available at: https://www.researchgate.net/profile/Stephen_Lawrie/publication/268202484_NeuroGrid_Collaborative_Neuroscience_via_Grid_Computing/links/5469cf060cf20dedafd10822.pdf

Gibaud, B., Kassel, G., Dojat, M., Batrancourt, B., Michel, F., Gaignard, A., et al. (2011). "NeuroLOG: sharing neuroimaging data using an ontology-based federated approach," in *Proceedings of American Medical Informatics Association, October 2011* (Washington, DC). Available at: https://hal.archives-ouvertes.fr/hal-00683087

Goldstein, L. B., Bertels, C., and Davis, J. N. (1989). Interrater reliability of the NIH stroke scale. *Arch. Neurol.* 46, 660–662. doi:10.1001/archneur.1989.00520420080026

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8:I1. doi:10.1186/1752-0509-8-S2-I1

González, D., Carpenter, T., van Hemert, J. I., and Wardlaw, J. (2010). An open source toolkit for medical imaging de-identification. *Eur. Radiol.* 20, 1896–1904. doi:10.1007/s00330-010-1745-3

Hanser, S., Boeker, M., Kumpf, K., and Schulz, S. (2007). "Design of an ontology on cerebral aneurysms: representing the conceptual space of the@ neurIST project. Medinfo 2007," in *Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems* (Amsterdam: IOS Press).

Heath, T., and Bizer, C. (2011). "Linked data: evolving the web into a global data space," in *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 1, 1–136. Available at: http://seco.cs.aalto.fi/u/jwtuomin/svn/secoweb/public_html/publications/2012/hyvonen-ch-book-2012.pdf

Hernández, M., Piper, R. J., Wang, X., Deary, I. J., and Wardlaw, J. M. (2013). Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review. *J. Magn. Reson. Imag.* 38, 774–785. doi:10.1002/jmri.24047

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imag.* 27, 685–691. doi:10.1002/jmri.21049

Job, D. E., Dickie, D. A., Rodriguez, D., Robson, A., Danso, S., Pernet, C., et al. (2016). A brain imaging repository of normal structural MRI across the life course: brain images of normal subjects (BRAINS). *Neuroimage*. doi:10.1016/j.neuroimage.2016.01.027

Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., et al. (2008). A National Human Neuroimaging Collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172. doi:10.1109/TITB.2008.917893

Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G. M., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi:10.1016/j.neuroimage.2013.05.094

Keator, D. B., Wei, D., Gadde, S., Bockholt, H. J., Grethe, J. S., Marcus, D., et al. (2009). Derived data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid. *Front. Neuroinformatics* 3:30. doi:10.3389/neuro.11.030.2009

Kim, B. J., Han, M.-K., Park, T. H., Park, S.-S., Lee, K. B., Lee, B.-C., et al. (2014). Current status of acute stroke management in Korea: a report on a multicenter, comprehensive acute stroke registry. *Int. J. Stroke* 9, 514–518. doi:10.1111/ijs.12199

Laird, A. R., Eickhoff, S. B., Fox, P. M., Uecker, A. M., Ray, K. L., Saenz, J. J., et al. (2011). The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Res. Notes* 4:349. doi:10.1186/1756-0500-4-349

Laney, D. (2001). "3D data management: controlling data volume, velocity and variety," in *META Group Research Note*, Vol. 6, 70. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Larson, S. D., and Martone, M. (2013). NeuroLex.org: an online framework for neuroscience knowledge. *Front. Neuroinformatics* 7:18. doi:10.3389/fninf.2013.00018

Lees, K. R., Bath, P. M. W., Schellinger, P. D., Kerr, D. M., Fulton, R., Hacke, W. D., et al. (2012). Contemporary outcome measures in acute stroke research: choice of primary outcome measure. *Stroke* 43, 1163–1170. doi:10.1161/STROKEAHA.111.641423

Lindley, R. I., Wardlaw, J. M., Whiteley, W. N., Cohen, G., Blackwell, L., Murray, G. D., et al. (2015). Alteplase for acute ischemic stroke: outcomes by clinically important subgroups in the Third International Stroke Trial. *Stroke* 46, 746–756. doi:10.1161/STROKEAHA.114.006573

MacKenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *Neuroimage* 42, 178–195. doi:10.1016/j.neuroimage.2008.04.186

Maillard, P., Delcroix, N., Crivello, F., Dufouil, C., Gicquel, S., Joliot, M., et al. (2008). An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology* 50, 31–42. doi:10.1007/s00234-007-0312-3

Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit. *Neuroinformatics* 5, 11–33. doi:10.1385/NI:5:1:11

Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). Making data sharing work: the FCP/INDI experience. *Neuroimage* 82, 683–691. doi:10.1016/j.neuroimage.2012.10.064

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2016). Best practices in data analysis and sharing in neuroimaging using MRI. doi:10.1101/054262

Pilat, D., and Fukasaku, Y. (2007). OECD principles and guidelines for access to research data from public funding. *Data Sci. J.* 6, 4–11. doi:10.2481/dsj.6.OD4

Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi:10.1038/nn.3818

Rector, A., and Rogers, J. (2006). *Ontological and Practical Issues in Using a Description Logic to Represent Medical Concept Systems: Experience from GALEN. Reasoning Web 2nd International Summer School* (Lisbon: Springer), 197–231.

Sandercock, P., Lindley, R., Wardlaw, J., Dennis, M., Lewis, S., Venables, G., et al. (2008). The third international stroke trial (IST-3) of thrombolysis for acute ischaemic strokev. *Trials* 9, 1–17. doi:10.1186/1745-6215-9-37

Sandercock, P., Wardlaw, J. M., Lindley, R. I., Dennis, M., Cohen, G., Murray, G., et al. (2012). The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *Lancet* 379, 2352–2363. doi:10.1016/S0140-6736(12)60768-5

Seghier, M. L., Patel, E., Prejawa, S., Ramsden, S., Selmer, A., Lim, L., et al. (2016). The PLORAS database: a data repository for predicting language outcome and recovery after stroke. *Neuroimage* 124, 1208–1212. doi:10.1016/j.neuroimage.2015.03.083

Smith, B., Arabandi, S., Brochhausen, M., Calhoun, M., Ciccarese, S., Doyle, B., et al. (2015). Biomedical imaging ontologies: a survey and proposal for future work. *J. Pathol. Inform.* 6, 37. doi:10.4103/2153-3539.159214

The ENOS Trial Investigators. (2006). Glyceryl trinitrate vs. control, and continuing vs. stopping temporarily prior antihypertensive therapy, in acute stroke: rationale and design of the efficacy of nitric oxide in stroke (ENOS) trial (ISRCTN99414122). *Int. J. Stroke* 1, 245–249. doi:10.1111/j.1747-4949.2006.00059.x

The ENOS Trial Investigators. (2015). Efficacy of nitric oxide, with or without continuing antihypertensive treatment, for management of high blood pressure in acute stroke (ENOS): a partial-factorial randomised controlled trial. *Lancet* 385, 617–628. doi:10.1016/S0140-6736(14)61121-1

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: an overview. *Neuroimage* 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041

Van Horn, J. D., and Toga, A. W. (2014). Human neuroimaging as a "Big Data" science. *Brain Imaging Behav.* 8, 323–331. doi:10.1016/j.neuroimage.2013.05.041

Walport, M., and Brest, P. (2011). Sharing research data to improve public health. *Lancet* 377, 537–539. doi:10.1016/S0140-6736(10)62234-9

Wang, F., Lee, R., Zhang, X., and Saltz, J. (2011). "Towards building high performance medical image management system for clinical trials," in *Proceedings of SPIE 7967, Medical Imaging 2011: Advanced PACS-based Imaging Informatics and Therapeutic Applications*. Lake Buena Vista, FL.

Wang, W., and Krishnan, E. (2014). Big data and clinicians: a review on the state of the science. *JMIR Med Inform* 2, e1. doi:10.2196/medinform.2913

Warach, S. J., Luby, M., Albers, G. W., Bammer, R., Bivard, A., Campbell, B. C. V., et al. (2016). Acute stroke imaging research roadmap III imaging selection and outcomes in acute stroke reperfusion clinical trials: consensus recommendations and further research priorities. *Stroke* 47, 1389–1398. doi:10.1161/STROKEAHA.115.012364

Wardlaw, J., Bath, P., Sandercock, P., Perry, D., Palmer, J., Watson, G., et al. (2007). The NeuroGrid stroke exemplar clinical trial protocol. *Int. J. Stroke* 2, 63–69. doi:10.1111/j.1747-4949.2007.00092.x

Wardlaw, J. M., Doubal, F., Armitage, P., Chappell, F., Carpenter, T., Muñoz Maniega, S., et al. (2009). Lacunar stroke is associated with diffuse blood–brain barrier dysfunction. *Ann. Neurol.* 65, 194–202. doi:10.1002/ana.21549

Wardlaw, J. M., and Mielke, O. (2005). Early signs of brain infarction at CT: observer reliability and outcome after thrombolytic treatment – systematic review. *Radiology* 235, 444–453. doi:10.1148/radiol.2352040262

Wardlaw, J. M., Muir, K. W., Macleod, M.-J., Weir, C., McVerry, F., Carpenter, T., et al. (2013). Clinical relevance and practical implications of trials of perfusion and angiographic imaging in patients with acute ischaemic stroke: a multicentre cohort imaging study. *J. Neurol. Neurosurg. Psychiatry* 84, 1001–1007. doi:10.1136/jnnp-2012-304807

Westra, B. L., Latimer, G. E., Matney, S. A., Park, J. I., Sensmeier, J., Simpson, R. L., et al. (2015). A national action plan for sharable and comparable nursing data to support practice and translational research for transforming health care. *J. Am. Med. Inform. Assoc.* 22, 600–607. doi:10.1093/jamia/ocu011

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer* 25, 38–49. doi:10.1109/2.121508

Wintermark, M., Albers, G. W., Broderick, J. P., Demchuk, A. M., Fiebach, J. B., Fiehler, J., et al. (2013). Acute stroke imaging research roadmap II. *Stroke* 44, 2628–2639. doi:10.1161/STROKEAHA.113.002015

Conflict of Interest Statement: The authors are aware of no conflict of interest that might bias the work presented here. Our funding sources had no involvement in this work.